

A human benchmark for automatic speaker recognition

Milou van Dijk¹, Rosemary Orr¹, David van der Vloed² and David A. van Leeuwen^{2,3}

¹University College Utrecht, The Netherlands

²Netherlands Forensic Institute, The Hague

³Radboud University Nijmegen, The Netherlands

Abstract

Automatic Speaker Recognition has a potential to be used in Forensic Speaker Comparison. For the latter, forensic scientists agree that presentation of the comparison to court should be in terms of a calibrated likelihood ratio. In recent years the field of automatic speaker recognition has made significant progress in the analysis, evaluation and calibration of likelihood ratios. In this paper we investigate if speaker comparison by humans can be carried out using the same framework. For this, we use the US National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation 2010 material to measure the performance and calibrate the speaker comparison opinions of human subjects. Because empirical calibration needs a large collection of trials and a human judgment takes a substantial effort, the analysis is carried out for a collection of 40 subjects. From NIST SRE 2010 a subset of 1280 speaker comparison trials are selected. The selection is made using the scores from a state-of-the-art speaker recognition system, such that 1) the trials are representative of the overall performance, in terms of difficulty of the comparisons for the automatic system, and 2) they can be analyzed in three distinct classes ‘hard,’ ‘representative’ and ‘easy.’ Results show that this classification extends to the performance of the human collective, with an Equal Error Rate of 45 %, 25 % and 13 % respectively. Further, the overall human results can be calibrated using ROC convex hull analysis to show a nice linear relation between the 10-level similarity response and a log-likelihood-ratio scale.

1. Introduction

With the performance of automatic speaker recognition systems steadily increasing, in part driven by evaluation campaigns such as the NIST Speaker Recognition Evaluations and the various JHU and BOSARIS workshops, and commercial systems becoming readily available, the use of automatic speaker recognition systems for forensic speaker comparison purposes becomes viable. In both automatic speaker recognition and forensic speaker comparison, the task is to compute the likelihood ratio that, given two speech segments, these originate from the same speaker or from different speakers. In formula form, the

likelihood ratio r is

$$r = \frac{P(\text{speech segments} \mid H_p)}{P(\text{speech segments} \mid H_d)}, \quad (1)$$

where $H_{p,d}$ are the *prosecutor's* and *defense* hypotheses, stating that the speech segments are produced by the same, or different, speakers, respectively. In automatic speaker recognition, the likelihood ratio can be used to make an optimal Bayes' decision given a cost function and a prior [1], whereas in forensic speaker comparison this can be used to express the weight of evidence in court [2].

In many countries, forensic speaker comparison is still carried out exclusively by human experts [3, 4], but in some countries forensic examiners are beginning to use automatic speaker recognition in certain cases [4, 5]. An argument, on the one hand, to be reluctant to use automatic systems in a forensic case is that the speech style and electro-acoustical recording conditions of the trace (the evidence) is often quite different from the reference recording in the speaker comparison, and no explicit performance characteristics of the system under those conditions are known. On the other hand, the reports of forensic speaker comparisons are seldom explicit in computation of the likelihood ratio for acoustic-phonetic or linguistic features that are marked as similar between the questioned recording and the reference [6].

The methods for computing likelihood ratios in the automatic speaker recognition domain and in the forensic speaker comparison domain are quite different. In the former, the approach is empirical and the raw discriminative scores of a system are taken as uncalibrated scores, and using a large collection of supervised trials (same-speaker and different-speaker comparisons) an empirical score-to-likelihood-ratio transformation is determined. So far, one of the most robust and effective score-to-likelihood ratio functions has been an affine transformation of the score s

$$\ell \equiv \log r = as + b, \quad (2)$$

effectively scaling the score by a and shifting it with b such that the resulting log-likelihood-ratio has good, probabilistically well interpretable, properties. In the latter, manually or semi-automatically obtained continuous features are directly modeled in same-speaker and

different-speaker distributions [7], or, in the case of discrete features, population frequencies can be used to compute the likelihood ratio, similar to how this happens with DNA. But often, an opinion is formulated where the similarity between the segments is expressed using a “verbal scale” [3,4,6,8]. How such a verbal scale maps to likelihood ratios is, however, a subject of debate. [9, 10]

If we want to get a better insight in how the automatic and the human methods compare we should probably let one do the task of the other, and see what the performance is. One way of doing this is by doing a “human benchmark”: giving a human exactly the same task as the system, and evaluate the performance in the same way. For automatic speaker recognition, such an experiment has been carried out by Schmidt-Nielsen and Crystal [11], and in the NIST Human Assisted Speaker Recognition (HASR) evaluations [12–14]. What the exact ‘human method’ is, is not so important for the performance evaluation and calibration method set forward in this paper; it could be a detailed acoustic-phonetic analysis as performed by forensic experts [3, 4], or holistic acoustic impressions as carried out in this study and others [11, 13, 14].

In this paper, we carry out a similar experiment to Schmidt-Nielsen in a somewhat different setting, and with the goal to investigate a method for determining a score-to-likelihood ratio mapping for human speaker comparison. This is in a way similar to the approach of ATVS-UAM to NIST HASR 2010 [13], where this mapping was taken a linear function. In this approach we study the shape of the mapping and the range of the resulting likelihood ratios.

2. Experimental design and details

With humans being limited in the amount of trials they can perform, we have designed an experiment with several goals in mind. The first is that the overall task should be of the same level of difficulty as the system test of NIST SRE 2010. Secondly, we want to study if trials that are hard for a system are also hard for the humans and vice versa. Finally, we want to find a relationship between a verbal scale of similarity and the likelihood ratio.

For empirical performance measurement and calibration we need many trials. Because a single trial takes a human subject a considerable time to complete—the subject should at least listen once to the segments—we decided to determine the performance of human subjects as a whole, effectively integrating out between-subject performance variation. Thus, every subject was exposed to their own set of trials, and the analysis is typically carried out over all subjects.

2.1. Trial selection

We used a different trial selection algorithm from what was used in HASR1 in NIST SRE 2010 [12], where pairs of similar speakers were sought for non-target trials and dissimilar target trials using a combination of machine and human judgments, leading to a set of very hard trials. In order to have a range of difficulties in this experiment we selected the trials as follows. We used speech material from the NIST 2010 SRE, telephone-telephone male English condition, a.k.a. ‘det 5’. Our Radboud University Nijmegen (RUN) automatic speaker recognition system [15] computed scores for all trials, i.e., the entire score matrix of train vs. test segments. Then we self-calibrated the scores using logistic regression, i.e., a linear transformation of the log-likelihood-ratio (LLR) scores (2) optimizing a cross-entropy objective function on the test data itself. The log-likelihood-ratio score distributions after calibration are shown in Fig. 1. We then selected trials from three regions, corresponding to difficulty classes: 1) around $\ell = 0$: these trials are “hard,” the recognizer cannot separate targets and non-targets; 2) around the modes in the distribution: these trials are “representative”; 3) high target and low non-target scores: these trials can be considered “easy.” These three classes are indicated as shaded bars in Fig. 1. The total amount of trials in each class was 160, 960 and 160 respectively. The trials were further distributed over 40 subjects in such a way that each subject had 4, 24 and 4 trials from each class, respectively, with equal amounts of target and non-target trials per class. This distribution guaranteed that a) each subject is exposed to approximately the same level of difficulty, according to the system, b) the target priors for each subject are the same, c) the overall difficulty is similar to the complete test. Since bias, the tendency of some subjects to find speakers more different where others may find them more the same, has an effect on the calibration of the speaker comparison opinion, we stressed that the target priors of the trials were 50 %. This is different from the experiment by Schmidt-Nielsen, where the priors were only *approximately* 0.5, as we did not anticipate subjects to count their own decisions to match the given priors over the 32 trials. For any subject, trials were presented in random order.

2.2. Experimental interface

We used a similar experimental interface to what we had used in experiments in human language recognition [16], that had proved to be quite effective. The interface is shown in Fig. 2. Every trial is a comparison of two speech segments, where the task is to determine if the identity of the speaker is the same or not. For both “same” and “different”, five levels of confidence could be specified, named “very uncertain”, “uncertain”, “confident”, “very confident” and “certain”. This configuration is the same

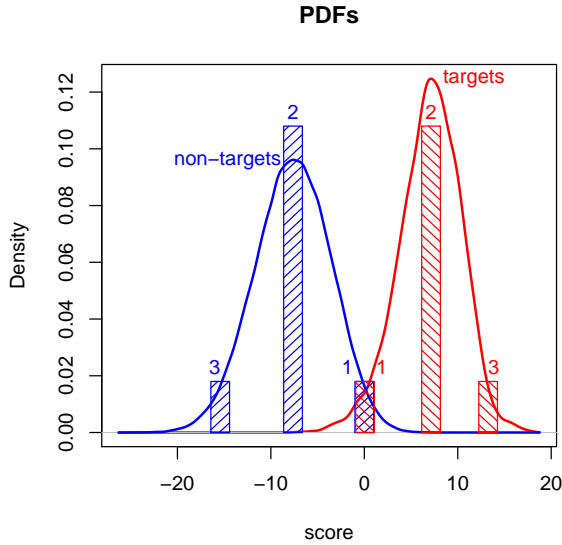


Figure 1: The Probability Density Functions (lines) for target (red) and non-target (blue) scores of the automatic speaker recognition system for NIST SRE 2010 ‘det-5’ male trials after self-calibration. The bars show the LLR score regions and relative quantities from which the trials for the human benchmark were drawn, with numbers indicating the difficulty class.

as in the Schmidt-Nielsen experiment, but the labels have somewhat different naming. The labels are different from the ‘verbal scale’ that is sometimes used in forensic evidence reporting [8], not only in wording (“... support for the prosecution hypothesis”) but also in the omission of an “indecision” option corresponding to a likelihood ratio (LR) of 1. Since the subjects were lay people w.r.t. forensic speaker comparison, we felt the wording in terms of a posterior probability was more intuitive, and is not incorrect given the explicit information about the prior. The subjects were in control of the playback of both segments, they could switch from one to another, and pause, at will. In order to maximize exposure to phonemic variability, playback of a segment would continue where it had been stopped the previous time. We did not provide the subjects with further control over the position of playback. A trial ended when any of the ten response buttons along the “certain: same”–“certain: different” scale was pressed. The interface did not allow for corrections after decisions. For comparison, in the experiment by Schmidt-Nielsen [11], a stimulus was presented as a test-target-test sequence of 3×3 s, paced by the subjects. Trials were presented in blocks of 20 with the same target speaker, with 2 min. training of the target speaker before each block. The ten response buttons were ordered left to right, with *extremely certain* at the edges and *uncertain* in the middle, similar to our vertical lay-out (cf. Fig. 2).

Are samples A and B spoken by the same speaker or by a different speaker?

5: certain	Same
4: very confident	Same
3: confident	Same
2: uncertain	Same
1: very uncertain	Same

Sample A

Pause

Sample B

1: very uncertain	Different
2: uncertain	Different
3: confident	Different
4: very confident	Different
5: certain	Different

Note: 50% of the trials have the same speaker and 50% of the trials have a different speaker

Figure 2: The experimental interface shown to the subjects

2.3. Recruitment of subjects

Because the focus of this study is on the evaluation and calibration method for human speaker comparison, we used naïve subjects rather than forensic experts—a much more scarce resource—in order to obtain more trials, similar to [13] and [14]. Forty subjects were recruited from the student community of the *University College Utrecht*, an international liberal arts and sciences establishment. The language of communication at the college is English, and most subjects are non-native English speakers but actively embedded in an English speaking community. Recruitment was carried out through contemporary social media, and subjects were not paid for their efforts. The typical session duration was 30–45 minutes, with some subjects requiring much more time. None of the subjects reported hearing problems, but their hearing abilities were not explicitly tested.

2.4. Experimental details

Experiments were conducted in a reasonably quiet environment. The software was run in a Java virtual machine in a Linux virtual environment on a laptop PC. Audio was presented through high quality headphones at a comfortable listening level. Longer periods of silence (> 0.5 s) in the speech had been automatically removed in order to make the experiment more efficient. The subjects received a short introduction about the purpose of the experiment from the experiment leader, and further received instructions through information panels on the screen. One of the screens drew special attention to the fact that in 50% of the trials the speakers were, in fact, the same. This message was also always visible during the main

Table 1: Aggregate statistics for the responses, same/different versus response score. Scores are sorted from “certain: different” to “certain: same.”

trial	-4.5	-3.5	-2.5	-1.5	-0.5	0.5	1.5	2.5	3.5	4.5
diff	152	117	152	74	11	14	40	43	18	19
same	36	26	70	56	22	15	78	138	118	81

experiment (cf. Fig. 2). After this information, six otherwise unused trials were presented as training/habituation. No feedback towards the decision was given in the habituation period, but the trials were chosen according to the easiest selection criteria.

After the experiment of 32 trials, subjects received immediate feedback about their performance in terms of an Equal Error Rate (EER).

3. Results

3.1. Overall performance

For analysis of the results, the response buttons are represented as scores, from +4.5 for “certain: same” to -4.5 for “certain: different”. The aggregate response statistics are tabulated in Table 1. The overall results can best be summarized in a Receiver Operating Characteristic (ROC), as in Fig. 3. This is a graph showing the trade off between the probabilities of false alarms (false positives, P_{FA}) and misses (false negatives, P_{miss}) if a threshold t for forcing decisions would have been ‘between the response buttons,’ effectively at $t = -5, -4, \dots, 5$. The circles correspond to these thresholds, and the line segments to response buttons. Because we have plotted the convex hull (CH) of the ROC, which has a special minimum-cost interpretation [17], some segments actually correspond to a group of adjacent buttons.

An often reported summary of the overall performance is the Equal Error Rate $E_{=}$, which we define as the point where $P_{FA} = P_{miss}$ on the ROC-CH. For the overall data $E_{=} = 26.5\%$. We can compute $E_{=}$ for the different subsets of the data, namely the difficulty classes 1–3 discussed in Section 2.1. From hard to easy, the results are $E_{=} = 44.8\%, 25.5\%, 13.2\%$.

We can use the ROC-CH to compute what the optimal log-likelihood-ratio score is corresponding to the buttons, assuming we can treat all subjects as a single ‘system.’ This implicitly assumes that subjects share the same ‘calibration,’ i.e., that one person means more-or-less the same with “confident” as the next. The optimal likelihood that can be associated with the subject’s judgment is just the negative slope of the corresponding ROC-CH line segment. Optimal in this sense means restricting the score-to-likelihood function to be a monotonously increasing function. The result of this operation is plotted in Fig. 4, as the heavy black line.

The likelihood ratio can also be computed by taking

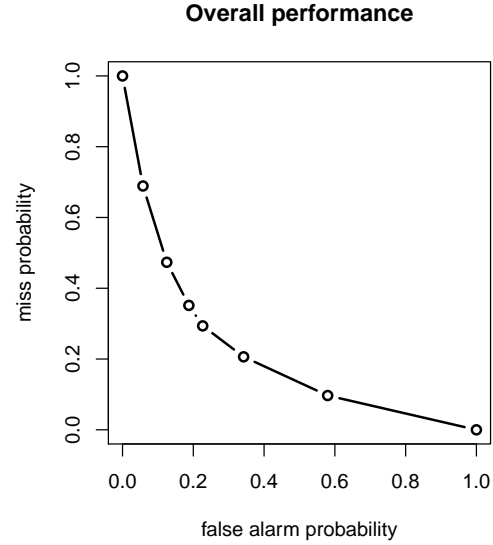


Figure 3: The overall performance, as Receiver Operating Characteristic, using the convex hull.

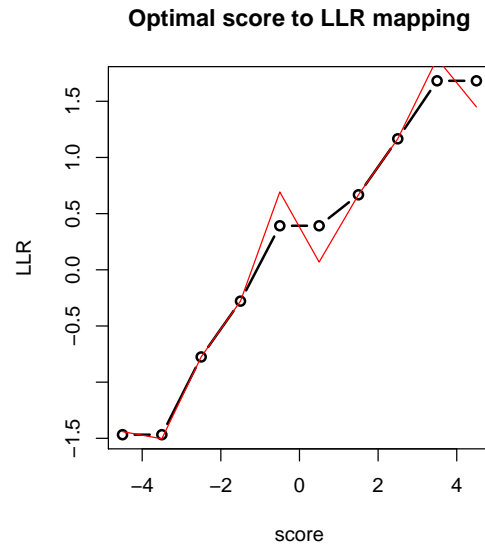


Figure 4: The optimal score-to-log-likelihood ratio function for the response of the human subjects (heavy line), and the LLRs found from ML ratios.

the Maximum Likelihood (ML) estimates of $P(s | H_p)$ and $P(s | H_d)$. From Table 1 we would, e.g., find for the score -0.5 (very uncertain: different) the ratio $\frac{22}{640} / \frac{11}{640} = \frac{22}{11}$, resulting in a log-likelihood ratio of $\log 2$. The values of such a computation are shown as the red, thin line in Fig. 4. It follows our earlier optimal LLR curve, but is not constrained by monotonicity. This way of computing LLRs directly from the probability density functions (PDFs) observed in development data sets [18] is popular in the forensic science community, but we feel that it has some undesirable properties. For one thing, it can associate a *higher* LR with a *lower* score, as can be observed from Fig. 4 at scores near -4 , 0 and 4 , but it can also lead to arbitrarily high and fluctuating LR’s if smoothing parameters that are needed to estimate the PDF for continuous features are chosen badly. The ROC-CH method is more robust in relation to these issues. Note that the ROC-CH method gives exactly the same LLRs as the PAV algorithm [1] does.

3.2. Per-subject calibration

As indicated above, the overall performance is expected to be a bit pessimistic because mis-calibration between subjects will lead to worse performance. We can try to compensate for this by calibrating the individual subject’s responses using their own performance characteristic, and then pooling their calibrated scores. As a first, cheating, experiment, we use all 32 responses per subject to compute this subjects optimal score-to-likelihood-ratio mapping. This is very similar to the “likelihood-ratio” score combination method used in [11] to combine responses from different listeners,¹ although we use the ROC-CH derived LR (cf. the black line in Fig. 4) and they use the ML method (red line). In order to limit the magnitude of the LLRs, which can easily become $\pm\infty$ for some trials, we used ‘Laplace’s rule of succession’ [1], which effectively adds additional scores of $+\infty$ and $-\infty$ to target and non-target scores before calibration, to cater for scores potentially unobserved in training. The resulting ROC is indicated in Fig. 5 in black, and it shows a much lower overall $E_{\pm} = 23.0\%$. This way of calibrating each subject individually really is “cheating,” because the information of the true hypothesis is used for each trial, albeit in a constrained way. If this monotonicity constraint were removed, such cheating would lead to LLRs of $\pm\infty$, giving rise to no errors.

A better approach to calibrating individual subjects is to use a cross-validation method. We used the chronological first half of each subject’s trials to compute an optimal score-to-LLR transformation, and applied this to the second half of their scores. In order to obtain the same number of scores as before we also reversed this

¹In [11], this was used to combine responses from subjects for the *same* trial, where we do it to pool different trials, but the idea is the same.

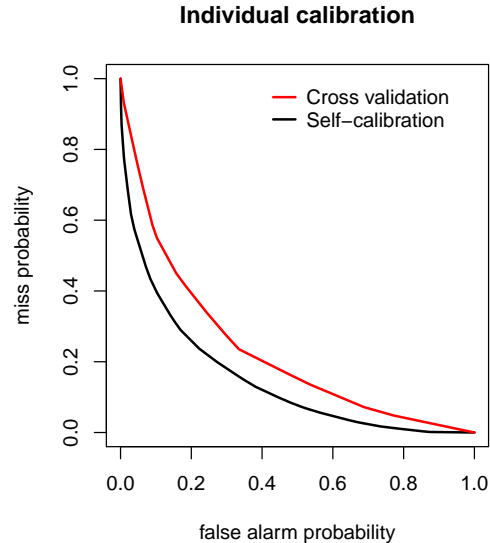


Figure 5: ROC after calibration of individual subject’s responses.

operation, i.e., calibrating on the second half and applying this to the first, resulting in a 2-fold cross-validation setup. The results of this individual calibration is shown in red in Fig. 5, which has a lower discrimination performance than the cheating experiment shown in black, with $E_{\pm} = 28.8\%$. This is not better than the original, uncalibrated, pooled scores, with $E_{\pm} = 26.5\%$. Apparently, the 16 scores available per subject are not enough to calibrate an individual. In [13], a similar effect (individual vs. pooled) was observed w.r.t calibration performance.

4. Discussion and Analysis

The overall discrimination performance of the human subjects, evaluated as if the judgments are from a single, consistent, system for the SRE 2010 data in our experimental conditions can be summarized as $E_{\pm} = 26\%$. This may appear very high, at first, but we have to point out the caveat that these subjects are naïve w.r.t. the task, and predominantly not native speakers of English. Moreover, the subjects did not use the full content of the speech files, but made decisions after having listened to part of the files. We recorded the exact button-press times, so that we can analyze the average time a subject was exposed to speech. For the overall set, this was 17.7 s per speech segment, ranging per subject from 8.8–34.2 s. Only a small correlation effect ($-0.4\%/s$, $p = 0.02$) could be found between a per-subject E_{\pm} and this listening time. More interesting perhaps is that the average duration measured over trial difficulty classes (cf. Section 2.1) drops as 18.8, 17.8, 15.6 s with decreasing difficulty, showing that the easier trials took less effort.

If we want to compare these results to automatic sys-

Table 2: RUN Automatic Speaker Recognition system performance as a function of duration of the speech segments.

duration (s)	5	10	20	40	80
$E_{=}$ (%)	23.3	13.2	6.5	4.4	3.4

tem performance, we have to be very careful. First of all, system performance increases steadily over time, and specifically for this data, because researchers improve their systems using this data as evaluation material. Further, systems have access to the full utterances, and hence use more information per trial, and it is not trivial to change the experimental set-up to allow human subjects to utilize the same amount of information. Probably a set-up with detailed play-back control and spectrographic tools, much as the forensic speaker comparison examiner has, would come closer to these goals but require much more effort on behalf of the subject, making an experiment at this scale (1280 trials) almost impossible. Please note that the detailed Human Assisted Speaker Recognition evaluations of NIST SRE 2010 [12] and 2012 only contained 15 and 20 trials, respectively, and required substantial effort from participants. With these restrictions in mind, we can compare the results on the same trials to the RUN system², which is not the best available system but probably has not yet been over-tuned to this data. We used several data sets where the utterances have been truncated in duration [19]. The discrimination performance results are in Table 2, from which we can conclude that the naïve human performs comparably to the RUN system using about 5 seconds per segment where the humans use 18 seconds.

Related to the relatively low discrimination performance, we find that the range of LLRs is fairly limited, roughly to a magnitude of 1.5 (cf. Fig. 4), corresponding to LR in the range 0.23–5.4. This means that these subject’s opinion of “certain” (the most extreme confidence available to them) correspond to LR of roughly $\frac{1}{4}$ and 5 for this data, which is certainly far away from the ‘verbal likelihood ratio scales’ as used in the literature, which can range from 10^{-4} to 10^4 [9]. One could argue that the data or the task simply is too difficult, and that there is the psychological effect that humans want to use the entire response scale for answers, and will scale accordingly. However, if we analyze the mean absolute score for the first and second halves of the trials per subject, we find the values 2.96 and 2.80 which shows, if anything, an opposite effect. One can also argue that a small trial set like this can’t produce any high magnitude LLR anyway, as by virtue of Laplace’s rule of succession values would be limited to $\sim \log 640 \approx 6.5$, in a case where all tar-

²Using a configuration of the system that was different from the system used in trial selection, so that scores for the selected trials are more evenly distributed than the shaded areas in Fig. 1

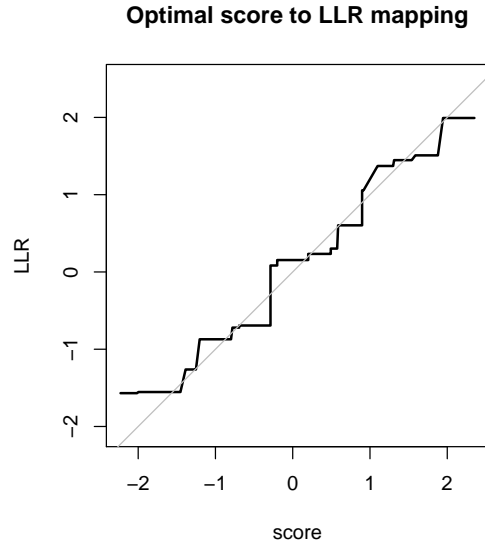


Figure 6: The score-to-LLR mapping after per-subject calibration, cross-validation experiment.

get trials are judged with the same, supportive, response, with no non-target trials with that response. The magnitude of LLRs corresponding to ‘certain’ suffer a bit from the mis-calibration between subjects. If we remove this by using the cross-validation calibration results from Section 3.2, as is shown in Fig. 6, the range of LR becomes a little wider, about $\frac{1}{5}$ –7.4. The finding of low magnitude LLR is consistent with [13] where calibrated LR not exceeding the range 0.1–10 were reported for human speaker comparison.

Despite the low magnitude LLRs observed in the original experiment in Fig. 4, it is interesting that the (raw) score-to-LLR function is fairly linear. No where in the briefing or the experimental protocol a logarithmic relation of the levels of certainty was suggested. Perhaps this is consistent with the interpretation of “weight of evidence,” which, for being an additive quantity functioning on a prototypical weighing scale of justice, must be expressed as log likelihood ratios [20]. In [13], this linear relation between the response value and the LLR was taken as an assumption in the calibration function. It appears we have found support for this assumption in this work.

The choice of ten response values, similar to [11], with a forced decision and visually not linear may have affected the distribution of responses. From Table 1 it can be seen that response values in the middle of the range, -0.5 and 0.5 receive less hits, and detailed analysis shows that this is true for all difficulty classes. We can only understand this as a psychological effect of evading ‘extreme’ responses, even though these middle values are not extreme in score value, but only in visual grouping. Such effects should be taken into account for

subsequent experiments, where we would advice to use a linear equidistant set of responses with a middle ground, “undecided” or “LR = 1.” Ramos [13] used a 7-point confidence scale with indeed a middle value indicating LR = 1. From Fig. 4, there is evidence that the correct interpretation of both “very uncertain” responses is $\ell \approx 0$, and it is probably better to allow explicitly for such a response.

5. Conclusions

The performance of the “human system,” whether internally calibrated or not, is not to be taken as representative for that of manual forensic speaker comparison: here, we work with lay listeners instead of experts, the exposure to the speech is very limited, the listeners are mostly non-native in the spoken language, long silences were removed, and the speech material is not taken from actual cases. However, the method of empirical calibration—as we are used to in automatic speaker recognition—should be possible to carry out with forensic experts. A practical problem, however, is the amount of effort that such an empirical calibration would take. If experts take, e.g., two weeks to form a well-founded opinion for speaker similarity for a single trial, a calibration at the scale of this experiment would take many person-years. Even if such an effort is taken, it is difficult to keep the ‘internal calibration’ of the expert constant over the entire period. Finally, the stimulus material must be made up of trials for which the true hypothesis is known, effectively ruling out real case material that includes the questioned recording. However, from ‘collateral’ case material it should be possible to generate trials where the hypothesis is known with negligible uncertainty. We would advocate that the proposed method of empirically calibrating opinions should somehow be carried out. Perhaps there are paradigms feasible in which the average time per trial is reduced, e.g., by grouping these for the same target speaker or by structurally including calibration trials in the standard operating procedure in forensic case work.

6. Acknowledgments

The research leading to these results has in part received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 238803.

7. References

- [1] N. Brümmer and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.
- [2] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, “Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2104–2115, September 2007.
- [3] T. Cambier-Langeveld, “Current methods in forensic speaker identification: Results of a collaborative exercise,” *The International Journal of Speech, Language and the Law*, vol. 14, no. 2, pp. 223–243, 2007.
- [4] E. Gold and P. French, “An international investigation of forensic speaker comparison practices,” in *The 17th International Congress of Phonetic Sciences (ICPhS)*, 2011, pp. 751–754.
- [5] H. J. Künzell, “Automatic speaker recognition with cross-language speech material,” *The International Journal of Speech, Language and the Law*, vol. 20, no. 1, pp. 21–44, 2013.
- [6] P. French *et al.*, “Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases,” *The International Journal of Speech, Language and the Law*, vol. 14, no. 1, pp. 137–144, 2007.
- [7] C. Aitken and D. Lucy, “Evaluation of trace evidence in the form of multivariate data,” *Applied Statistics*, pp. 109–122, 2004.
- [8] C. Champod and I. W. Evett, “Commentary on broeders (1999): ‘some observations on the use of probability scales in forensic identification’,” *Forensic Linguistics*, vol. 7, no. 2, pp. 238–243, 2000.
- [9] P. Rose, *Forensic Speaker Identification*. Taylor & Francis, 2002.
- [10] A. Nordgaard *et al.*, “Scale of conclusions for the value of evidence,” *Law, Probability and Risk*, vol. 11, no. 1, pp. 1–24, 2011.
- [11] A. Schmidt-Nielsen and T. H. Crystal, “Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data,” *Digital Signal Processing*, vol. 10, pp. 249–266, 2000.
- [12] C. Greenberg, A. Martin, L. Brandschain, J. Campbell, C. Cieri!, G. Doddington, and J. Godfrey, “Human assisted speaker recognition in nist sre10,” in *Proc. of Odyssey Speaker and Language Recognition Workshop*. ISCA, 2010, pp. 180–185.
- [13] D. Ramos, J. Franco-Pedroso, and J. Gonzalez-Rodriguez, “Calibration and weight of the evidence

- by human listeners. the ATVS-UAM submission to NIST human-aided speaker recognition,” in *Proc. ICASSP*. IEEE, 2011, pp. 5908–5911.
- [14] J. Kahn, N. Audibert, S. Rossato, and J.-F. Bonastre, “Speaker verification by inexperienced and experienced listeners vs. speaker verification system,” in *Proc. ICASSP*, 2011, pp. 5912–5915.
 - [15] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “The effect of noise on modern automatic speaker recognition systems,” in *Proc. ICASSP*. Kyoto: IEEE, March 2012.
 - [16] R. Orr and D. van Leeuwen, “A human benchmark for language recognition,” in *Proc. Interspeech*. Brighton: ISCA, September 2009, pp. 2175–2178.
 - [17] N. Brümmer, “Measuring, refining and calibrating speaker and language information extracted from speech,” Ph.D. dissertation, Stellenbosch University, 2010.
 - [18] D. Meuwly and A. Drygajlo, “Forensic speaker recognition based on a Bayesian framework and gaussian mixture modelling (GMM),” in *2001, a speaker Odyssey*, June 2001, pp. 145–150, crete.
 - [19] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, “Quality measure functions for calibration of speaker recognition system in various duration conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, Accepted for publication 2013.
 - [20] I. J. Good, *Bayesian Statistics 2*. Elsevier Science Publishers, 1985, ch. Weight of Evidence: A Survey, pp. 249–270.